Griffin JM Bur
SOC 361

Explaining income variation with a naïve measure of class and in-class inequality

I. Introduction

In this paper, I contribute to the social stratification literature — which typically uses theories of status, in Weber's sense, as its central explanatory variable — by constructing a scale of class, compatible with both Weber and Marx. I make a subsequent contribution to the literature by testing whether this measure helps explain variations in income. I hypothesize that my constructed measure of class can help explain variations in income when added to basic predictive models of income based on respondents' race, gender and education data.[1] In the course of that assessment, I also make some significant discoveries about the ways in which the effects of race on income are moderated by other variables. Below, I show my results, closely following the proposed paper sections on the rubric.

I.i. A new measure of class

While concrete measures of status or prestige have been developed empirically through scales such as the Nakao and Treas index, few reliable measures of class exist.[2] Though some notable hybrid measures of class and status exist, such as the Duncan socioeconomic index, one tricky feature of Duncan index-type measures is that the typical pay of the respondent's occupation is included in the scale. This means that one of the ways in which we might want to test the validity of a status or class construct — by gauging how well the construct explains variations in income — would be partially redundant.[3] So, I construct a measure of class below that does not include any data about the respondent's actual pay or the typical pay of their occupation (although some Likert-type items asking respondents to evaluate their relative financial situation are included). Instead, I largely rely on information about workplace hierarchies as well as respondents' reported feelings of worth and respect in the workplace to provide an

---

[1] This is a loose hypothesis; technically, given the way I phrase it, it would be true so long as the scale added at all to the R^2 of the model. This would be inappropriate for a journal article but since this paper's purpose is methodological, I believe it is functional for my purposes. I originally included a model using race, gender, education and occupational prestige for comparison but this proved to take up too much space.

[2] On the Nakao and Treas measure, see Michael Hout, Tom W. Smith and Peter V. Marsden, "Prestige and Socioeconomic Scores for the 2010 Census Codes", *GSS Methodological Report No. 124* (n.d.).

[3] On the construction of the Duncan index: Robert M. Hauser and John Robert Warren, "Socioeconomic Indexes for Occupations", *Social Methodology* 27 (1997): 177-298. While the respondent's actual income and the typical income of her or his occupation are, of course, not necessarily the same value, this does mean that the Duncan index is not a completely "independent" measure of social class — at least in my opinion.

estimation of individual's class position. I refer to this scale as "proxy class" because, in principle, it would be possible to use survey data to actually measure the class status of individuals if more information were available; this is a project far outside of the boundaries of this class, however.[4]

I.ii. Explaining income

I test the power of my measure by regressing respondents' imputed incomes[5] on their class scores. While the concept "class" exists at a higher level of abstraction income, and is often incorrectly conflated with income, class should nevertheless be linearly associated with income and should be able to explain some of the variation in income if combined with appropriate control variables. My basic hypothesis is that the measure of class I have constructed will help variations in the respondent's income when added to a simple predictive model using race, gender and education information.

I.iii. Some secondary findings about my control variables

In the course of testing the explanatory power of class, I also find that race categories have surprisingly little power to explain individual or family income when class (and, for that matter, education and/or education) is included in the model. I show that this provides support for theories about the mediating role of class and education for the effects of race. That is, while race remains a significant line of social fracture, its direct effects on income appear to be suppressed by class.[6]

II. Data and variables

In this paper I use the General Social Survey (GSS) 2012 dataset, collected by the National Opinion Research Center (NORC) which is a non-profit, long-standing national survey center affiliated with the University of Chicago.[7] The survey was conducted with what the GSS authors call "full probability sampling"; with a 71.4 percent response rate, they were able to receive responses from 1,974 US adults. My analytical sample after accounting for missing values was significantly smaller — 794 cases — but still large enough to generate analytically-robust results. The source of the missing data is

[4] I am thinking here of an example such as E.O. Wright, *Classes* (New York: Verso, 1985).
[5] I impute income using the unpublished "topcat" code generously provided by John A. Logan.
[6] Nevertheless, the "mean class score" for Black respondents was significantly lower than that of whites. An individual's "race" makes her more likely to get a certain score on the class variable, but her race does not have independent effects once her class score is controlled for.
[7] This description is a paraphrase of their own self-description on their website.

simply the fact that I use many questions from the 2004 topical module called "workplace conflict" which was only completed by about ⅔ of respondents.[8] A total of 795 respondents marked "iap", or inapplicable, on these questions; this means that there only 385 cases were "genuinely" dropped from a total of 1,974 cases due to random missing data.

I will begin my description of variables with those that come in the dataset or which I lightly recoded which serve as control variables: gender/sex, race/ethnicity, and education. I have decided to forego a table of the variables since only one is a continuous variable (education) and even that variable is eventually converted into a discontinuous variable; instead, each subsection below provides summary statistics about the variable in question. I then move on to describing my major explanatory variable (class). I finish up with a discussion of my response variable (income).[9]

II.i Gender/sex

The GSS category "sex" is given as a binary variable where "1" is male and "2" is female. As is typical for survey data, women are overrepresented: 1,089 respondents identify as women, and 885 identify as men. There were no missing values. I created a new variable called "gender" where "1" refers to women and "0" to men, so that men were the reference category. I use the term "gender" because sex has come to refer strictly to biological endowments at birth in the social sciences literature, whereas gender refers to the way an individual identifies and is "socially coded". The GSS codebook does not provide the exact wording of the question, but respondents will presumably select their gender rather than their sex. In my analytical sample, there were 452 men and 468 women; about 49 percent of respondents were men and 51 percent were women. This is not too far off from the full sample, where 45 percent of respondents (885 people) were men, and 55 percent (1, 089 people) were women.

II.ii Race

The GSS categories about race are somewhat tricky to handle. There are only three possible values for the question "What race do you consider yourself?" on the GSS: white, black and other (see Figure 1 below).

Figure 1. Distribution of respondents by GSS race categories.

[8] GSS Codebook 1618

[9] Please note that some data manipulation is treated in more detail in the appropriate sections (for example, the discussion of logged income is discussed in the section on nonlinearity).

	n	percent	cumulative
white	1477	74.82	74.82
black	301	15.25	90.07
other	196	9.93	100.00

Fortunately, the GSS does include a question about whether respondents identify as Hispanic or not, entirely separate from the variable "race" just described. In total, 1,970 respondents answered the "Hispanic" question; 1,701 said "no" (86 percent) and 269 said "yes" (14 percent). Unfortunately, but predictably, there was some overlap between respondents who answered "yes" on "Hispanic" and respondents who answered "yes" on any of the other three race dummies. Fortunately, the GSS does include a question about whether respondents identify as Hispanic or not, entirely separate from the variable "race" just described. In total, 1,970 respondents answered the "Hispanic" question; 1,701 said "no" (86 percent) and 269 said "yes" (14 percent). Unfortunately, but predictably, there was some overlap between respondents who answered "yes" on "Hispanic" and respondents who answered "yes" on any of the other three race dummies.

So, I performed the following transformations. First, based on the well-known theory that US racism functions on the basis of the "one-drop" rule for persons of African descent, I coded all Black Hispanics as Black, assuming that their African ancestry is more relevant to how they are "racialized" than their Hispanic ancestry. Second, I recoded white Hispanics so that anyone answering yes to both "white" and "Hispanic" was coded as white. Since "white" is much closer to being a race category than an ethnic category, and because the social fiction of biological race is more important for determining, in the context of race-domination, that person's life chances than that person's actual ethnic identity, it seems safe to assume that someone's ability to benefit from whiteness matters more, for my purposes, than the fact that, for example, that person's first language may be Spanish.

Finally, I re-coded all persons who answered "1" on both "other" and "Hispanic" to "Hispanic only". This was a simple choice to make. Since "other" is such a heterogeneous and unsatisfactory category, it is clearly preferable to use a more specific category where there is some reason to do so. There is still a possibility that some persons who answered both "other" and "Hispanic" may have also belong to a more relevant race-ethnicity categories — perhaps some Black Hispanics answered "other" to

indicate this — but there is no way of recovering this information without redoing the survey.

If this procedure is performed before the large dropping of cases I performed, the distribution is: white, 1309 cases (66 percent); Black, 301 cases (15 percent); Hispanic, 265 cases (13 percent), other, 99 cases (5 percent).[10] If the procedure is performed after the cases are dropped, the distribution is: white, 514 cases (65 percent); Black, 117 cases (15 percent); Hispanic, 125 cases (16 percent), other, 38 cases (5 percent). These two sample statistics are comparable, which is a good sign.

II.iii. Education

Education is the easiest category since it is ordinal and fairly uncontroversial. The question is asked in a fairly complicated way on the GSS, however, but the upshot is that respondents can essentially answer with any whole number between "0" and "20". There are two missing values on the GSS-provided category ("educ") so I dropped them, leaving 1972 cases. After that, in the full sample, the mean year of education completed was 13.5, the median value was 13, and the standard deviation was 3.1. Values ranged from 0 to 20. In the analytical sample, the mean year of education completed was 14.1, the median value was 14, the standard deviation was 3.01 and the values ranged from 1 to 20.

For the sake of capturing non-linear effects of education on income, I later broke up this variable into five dummies: "less than high school education" (0-11), "high school education" (12), "some college" (13-15), "BA" (16), and "postgrad" (20). I talk more about the non-linear effects of education in the section on non-linearity below.

II.iv Class

In order to construct a rough estimate of class, I decided to pull some variables that are very indirect measures of class that nonetheless hang together fairly well to construct a scale of class for use as an explanatory variable; Cronbach's alpha is 0.80. The questions primarily measure how satisfied the respondent is at work in different ways, although some "objective" questions that indirectly get at class are included to make sure that my scale does not simply measure "anger at work" — such as whether or not the respondent supervises others or is supervised by others at work.[11] So, as noted above, this is a

[10] Please note that the percents do not sum to 100 due to rounding error.
[11] I had foreseen this danger but I would also like to acknowledge that Joshua McAuliffe also diagnosed this potential foible in comments on a presentation of this paper given at UW-Madison on 2 May 2017.

proxy for class; most of the variables are subjective and thus rely on respondents' self-perception.

My scale can be thought of as measuring one very general phenomenon — social stratification — at two levels of abstraction: class (the higher and broader level) and hierarchy within classes (the lower and more continuous level). In a sense this is akin to the way in which the education variable simultaneously measures broad categories encompassing multiple values (such as "postgraduate degree" and "no high school degree") and more specific values within those categories (such as the exact years of education that non-high school graduates have). The difference is that my class variable is continuous; also, I have not broken up the scale into ranges corresponding to, say, workers and owners — this would be well outside the bounds of the paper and it is not immediately obvious where those cut-off points should be on the scale, anyways.

The items I put into my scale are meant to capture a wide variety of data that indirectly indicate someone's class position. The latent variable that the scale tries to measure is, as noted, 1) whether or not someone primarily makes their living from wages rather than profits and 2) relative privilege within those broad groups (which can be thought of as the probability of their moving into another class). The full set of variables is included as an appendix to the paper since it is rather long.

The scale has a range of values from -0.353 to 2, with a mean of 1.13, a median of 1.18, and a standard deviation of 0.41. In my regression analyses, I standardized the variable so that its mean was 5.2, its median was 5.3, and its range was 1.6 to 7.3.

II.vi Income

I use the "rincom06" category which asks respondents to place their personal income into "buckets" of unequal sizes. This is the most detailed data available on respondents' incomes (some other questions simply ask whether the respondent makes more or less than $25,000 a year). I used the "topcat" command to estimate midpoints for these buckets and then computed the natural log of this figure. In the full sample, the median using topcat was about $27,000; the mean was about $41,000. The standard deviation was $42,000. For 723 respondents, this question was inapplicable; 105 declined to answer. When I changed the order of my operations, and dropped non-responders on the workplace questions before "topcatting" income, there were only 113 missing values on the income question, meaning that of

the total 828 non-responders on income, 795 of them were also non-responders on the workplace question. Since most of the cases that are missing workplace values are missing income values, and vice versa, it seems clear that the great majority of missing values on both questions like mean that the individual is retired or permanently outside of the labor force. They are thus not really part of my population of interest. I do not provide summary statistics of the logged income variable since this is not a very intuitive measure, and since no predictions are made.

II.vii Correlations of variables

Below is a correlation table of my variables. Correlations which do not have an intuitive, substantial meaning such as that between the race of respondents, are not included.

Figure 2. Correlation of variables used in the analysis.

	Income	Education	White	Black	Hispanic	Other race	Class
Income	1	-	-	-	-	-	-
Education	0.43	1	-	-	-	-	-
White	0.10	0.20	1	-	-	-	-
Black	-0.08	-0.03	/	1	-	-	-
Hispanic	-0.12	-0.31	/	/	1	-	-
Other race	0.11	0.12	/	/	/	1	-
Class	0.21	0.16	0.10	-0.09	-0.05	0.01	1

III. Model specification and assessment

Below, in Table I, I estimate income using four initial models, beginning with a simple race-only model and then building up to the full model in equation (IV) below.[12] The models are specified as follows. I estimate Model I as:

$$Y_i = \hat{B}_0 + \hat{B}_1 D_1 + \hat{B}_2 D_2 + \hat{B}_3 D_3 \qquad \text{(I)}$$

where y = income in thousands of dollars, D_1 is a dummy for Black respondents, D_2 is a dummy for Hispanic respondents and D_3 is a dummy for "other" respondents (white are the reference category). The null hypothesis for this model is that there will be no relationship between any of the race dummies

[12] I show this side-by-side "build up" in order to demonstrate a suppressor effect, not because any of the three reduced models are particularly plausible.

and income; that is, H_0: $\hat{B}_1 = \hat{B}_2 = \hat{B}_3 = 0$. My alternative is that: $\hat{B}_1 < 0$, $\hat{B}_2 < 0$, $\hat{B}_3 =/= 0$. In other words, Black and Hispanic respondents will have less income on average than white respondents owing to the long history of race-domination in the United States. I also hypothesize that "other" respondents will have higher or lower income than whites; however, because of the heterogeneity of the category, I do not have any way of knowing which sub-groups "other" is drawn from. Perhaps this category is largely comprised of working-class refugees; perhaps it is comprised largely of relatively wealthy expatriates So, I will consider one-tailed *p*-values for the first two hypotheses and a two-tailed *p*-values for the third.

I estimate Model II as:

$$Y_i = \hat{B}_0 + \hat{B}_1 D_1 + \hat{B}_2 D_2 + \hat{B}_3 D_3 + \hat{B}_4 D_4 \qquad \text{(II)}$$

where D_4 is a dummy for gender (with men as the reference category). My model here is a "nested extension" of the first model. For this model, the null is once again that H_0: $\hat{B}_1 = \hat{B}_2 = \hat{B}_3 = \hat{B}_4 = 0$. My preferred alternative is H_a:: $\hat{B}_1 < 0$, $\hat{B}_2 < 0$, $\hat{B}_3 =/= 0$, $\hat{B}_4 < 0$. The coefficient for women is predicted to be negative because there is evidence that women have a subordinate position in the US.

I estimate Model III as:

$$Y_i = \hat{B}_0 + \hat{B}_1 D_1 + \hat{B}_2 D_2 + \hat{B}_3 D_3 \hat{B}_4 D_4 + \hat{B}_5 X_1 \qquad \text{(III)}$$

where X_1 is a continuous variable indicating respondent's years of education. I again have a nested extension, this time of Model II. My null is H_0: $\hat{B}_1 = \hat{B}_2 = \hat{B}_3 = \hat{B}_4 = \hat{B}_5 = 0$. My preferred alternative is H_a: $\hat{B}_1 < 0$, $\hat{B}_2 < 0$, $\hat{B}_3 =/= 0$, $\hat{B}_4 < 0$, $\hat{B}_5 > 0$. The effect of education on income is predicted to be positive because it transmits skills to individuals that allows them to bargain for higher wages.

I estimate Model IV as:

$$Y_i = \hat{B}_0 + \hat{B}_1 D_1 + \hat{B}_2 D_2 + \hat{B}_3 D_3 + \hat{B}_4 D_4 + \hat{B}_5 X_1 + \hat{B}_6 X_2 \qquad \text{(IV)}$$

where X_2 is a continuous scale indicating class (which is standardized). My hypothesis here is expressed as H_0: $\hat{B}_1 = \hat{B}_2 = \hat{B}_3 = \hat{B}_4 = \hat{B}_5 = 0$. My preferred alternative is H_a: $\hat{B}_1 < 0$, $\hat{B}_2 < 0$, $\hat{B}_3 =/= 0$, $\hat{B}_4 < 0$, $\hat{B}_5 > 0$, $\hat{B}_1 > 0$. I hypothesize that the effect of my class scale will be positive for the reasons laid out in my discussion of class and income above.

TABLE 1 *Estimated effects of race, gender, education and class on imputed incomes.*[13]

Variables				
	Model I	Model II	Model III	Model IV
Controls				
Black	**-10.99*** **	**-9.22** **	-5.05	-3.55
	(4.19)	**(4.11)**	(3.74)	(3.72)
Hispanic	**-14.99*** **	**-16.67*** **	-1.44	-1.05
	(4.08)	**(4.00)**	(3.81)	(3.77)
Other race	**17.63*** **	**16.63** **	9.96	10.47*
	(6.88)	**(6.74)**	(6.13)	(6.06)
Gender		**-17.40*** **	**-18.55*** **	**-18.13*** **
		(2.86)	**(2.59)**	**(2.56)**
Education			**5.90*** **	**5.61*** **
			(0.45)	**(0.45)**
Class effects				
Class (standardized)				**5.87*** **
				(1.30)
Observations	794	794	794	794
R-squared	0.034	0.077	0.241	0.261

Standard errors in parentheses.
Results are **bolded** when significant at $\alpha = 0.05$ *for the appropriate tail-test.*
* $p < 0.1$
** $p < 0.05$
*** $p < 0.01$

Source: GSS 2012

[13] For *all Stata* output, please see Stata files labeled "GJMBur361final".

I now interpret my initial results, beginning with Model I. As predicted, the Black and Hispanic dummy variables have *negative* coefficients which are statistically and substantively significant different than those of the reference category. Though it is not especially germane, note that the Hispanic and Black coefficients are not statistically different from one another (please see Stata output, Part III.i). The other race group is statistically different significantly from Black and Hispanic respondents and, interestingly, it is strongly positive; while I predicted a non-zero coefficient, such a large, positive coefficient was unexpected. Without knowing more about the composition of this category, however, I am hesitant to make too much of these results (especially given the large standard error). It is worth noting that at this point, however, very little of the variation in incomes is explained (three percent). I will delay commentary on causality until the final model is estimated, since my further findings complicate what — at the outset — appears to be a fairly straightforward story about the effects of interpersonal race-prejudice as well as suprapersonal race domination on income.

In Model II, my alternative hypothesis is again confirmed in full; each coefficient is significant at the $\alpha = 0.05$ level. The three race dummies' coefficients have not changed much in substantial terms; now, gender has a substantial negative coefficient (larger than any of the race dummies, in fact), as predicted. The Black and Hispanic coefficients again do not differ. While the amount of explained variation in incomes has increased in relative terms (more than doubling to seven percent), it is still low in absolute terms. I again hold off on detailed causal interpretation for now.

In Model III, the alternative hypothesis is only partially confirmed. Education has a substantial positive slope (especially once we account for the fact that educational achievement usually comes in multiple years) and gender still has a substantial negative slope. Interestingly, however, the effects of each of the three race variables has lost its significance. The *R*-squared of the model also significantly improves to 0.25. This suggests that the effect of race is mediated by the effect of education, perhaps through a kind of sorting mechanism whereby the educational outcomes of individuals (who are racialized at birth) depend on their race, and their income depends upon their education (this is explored later). A brief look at differential education outcomes by race (Figure 3 below) lends this intuition some initial credence.

In Model IV, the alternative hypothesis is again confirmed partially. The three race dummies remain insignificant; gender and education continue to have statistically-significant coefficients in the appropriate directions. The class variable (which is standardized) has, as predicted, a positive coefficient, although its inclusion adds little to the explanatory power of the model. The effect of a one standard deviation-increase or decrease in class is also significantly lower than a one standard deviation-increase in education, given that the coefficients above are similar but education has *not* been standardized like class (its coefficient would be 3.02 times larger if so).[14]

So far, it appears that Model IV is the preferred model, although some of the causal pathways implied still need to be investigated. In particular, it appears that the negative effects on income of the two "races" which are associated with marginalization in the US — non-white Hispanic and Black — are wholly mediated by education. It is also worth noting that if class is entered into Model II on its own, before and without education, its effect is about 40 percent larger than in Model IV (this is demonstrated below) — so education may suppress the effect of class to some degree.

Some interactions also need to be tested. At the outset I was interested in potential interactions between race and gender, and race and class, although these questions are perhaps now irrelevant. Some of my other proposed interactions — between education and gender, class and gender, and class and education — are still worth testing, however. The interaction of education and class might be justified with reference to non-linear effects of education: perhaps the degree to which education moderates the effect of class is greater at the BA level than at the high school diploma level. The interaction of education and gender also seems to make intuitive sense: given how large a mediating role education plays for race, perhaps the same is true for gender (i.e., perhaps the gender gap is smaller at higher levels of education). Finally, the interaction between class and gender also seems intuitively plausible: perhaps the effect of the gender gap is higher at the lower end of the class spectrum than at the higher end, or vice versa. The "general theory of intersectionality" which posits that different forms of social domination combine in unexpected, non-additive ways provides a broad theoretical grounding for these hypotheses.

[14] In Stata output III.iv, I generate this standardized education variable and perform an *F*-test with a null hypothesis that its coefficient equals that of the standard class variable — there is strong evidence to reject it.

IV. Non-linear specifications

I make two nonlinear modifications to the "full model" specified above. First, I convert the ordinal education variable into a series of five dummy variables: "less than high school education" (0 - 11 years), "high school education" (12 years), "some college" (13-15 years), "BA" (16 years) and "postgrad" (17-20 years). Model V below estimates this model with a less-than-high-school education as the reference category; Model VI, perhaps more realistically, uses a high school education as the reference category. In both cases, the estimates of individual coefficients were all significant at $\alpha = 0.05$.[15] When an F-test of the reduced model, where changes in Y were constrained to be the same across the dummy variables, was performed, there was strong evidence to reject the null hypothesis ($F = 9.10$), meaning that there was evidence for the non-linearity of education's effects.[16] Finally, the *R*-squared improves from 0.26 (Model IV) to 0.29 when education is "dummied up" (please see either Model V or Model VI in Table 2 on the next page). An initial interpretation of these results is that there is a fairly large financial benefit to completing a bachelor's, a relatively small incremental benefit for a post-graduate degree, a small penalty for not finishing high school, and no statistically-significant financial benefit for having only "some college". Given this last finding, I also estimated a model that collapsed the education categories (Model VII in Table 2 below) but it actually had a slightly worse fit so I retained the original model.

Second, I log income and regress it on natural units of my predictor variables; when I regress natural units of income on the predictions from this log-lin model, so that the two models may be compared, the log-lin model has a slightly higher *R*-squared (0.298) than the lin-lin model (0.290), so I use this model for the remainder of the paper; it appears below in Table 2 as Model VIII.[17] I also decide to follow the advice of John A. Logan and use the high school education dummy as the excluded category.

---

[15] Please note that in both models, the coefficient contrasting "less than high school education" and a high school diploma is significant at $\alpha = 0.10$ for a *two*-tailed test. But, since the direction of each coefficient is in the predicted direction — based on the hypothesis that more education means more money— it is significant for a one-tailed test.

[16] Please see section IV.i.a of the *.do* file / Stata output for this test.

[17] I do not show this converted output since it requires estimating *Y*-hats in Stata and regressing natural units of the outcome variable on these estimates; this method unfortunately collapses the different coefficients and variables by "plugging in" values, so that the final output appears as a bivariate regression — there is little learned by showing this output. I would like to thank Joshua McAuliffe for confirming my suspicions about this point.

V. Interactions

Initially, I argued for the possible presence of an interaction between 1) education and gender, 2) class and gender, and 3) class and education. I also argued for the possibility of an interaction between race and class as well as race and gender and race and education. Since race ended up being dropped from my model, I do not evaluate the latter three forms of interaction.

I first address the surprising non-significance of the first two interactions (see Table 2 below) and then discuss the more complicated third set of interactions. I should note that though Table 2 shows interactions for interactions between gender and the separate education dummies, interactions between gender and the single polytomous education variable were also insignificant (Stata output, Part V).

TABLE 2 *Three non-linear specifications of the gender, education and class model*

Variables				
	Model V	Model VI	Model VII	Model VIII (log-lin model)
Gender	**-17.02*** **	**-17.02*** **	**-16.88*** **	**-0.362*** **
	(2.50)	**(2.50)**	**(2.50)**	**(0.0632)**
Less than high school education		**-7.91***	**-9.71*** **	**-0.338** **
		(4.57)	**(1.27)**	**(0.115)**
High school education	**7.91***			
	(4.57)			
Some college	**11.56** **	3.65	/	**0.206** **
	(4.59)	(3.47)	/	**(0.0876)**
BA	**43.82*** **	**35.91*** **	**34.08*** **	**0.834*** **
	(4.91)	**(3.85)**	**(3.44)**	**(0.0972)**
Postgrad	**49.25*** **	**41.32*** **	**39.51*** **	**0.952*** **
	(4.80)	**(3.73)**	**(3.31)**	**(0.0942)**
Class	**5.57*** **	**5.57*** **	**5.59*** **	**0.124*** **
	(1.28)	**(1.28)**	**(1.28)**	**(0.115)**
Observations	797	797	797	797
R-squared	0.290	0.290	0.289	0.254

Standard errors in parentheses.
Results are **bolded** when significant at $\alpha = 0.05$ *for the appropriate tail-test.*
* $p < 0.1$
** $p < 0.05$
*** $p < 0.01$

Source: GSS 2012.

The non-interaction between education and gender, and class and gender, is somewhat surprising. On the one hand, it has become common sense in parts of the social sciences that different forms of social marginalization combine in complex ways; the eminent Black feminist scholar Patricia Hill Collins could already refer to "additive models of oppressions" as hopelessly outdated as early as 1990.[18] So, it is reasonable to intuit that there may be a statistical interaction between, say, gender and class; perhaps the effect of being working class is different from men than women, for example. On the other hand, however, it is worth taking a closer look at the exact meaning of the claim that "oppressions intersect" or "oppressions stack". It is tempting to apply such claims by simply modeling an "intersection" of oppressions as an interaction among variables, but it should be recalled that one of the central claims of theories such as those of Collins is that some forms of social domination are "*non*-economic". So, once one controls for an "economic"[19] factor, it is reasonable to assume that the economic effects of race and gender domination — "non-economic" forms of domination — would be muted or even suppressed, and there may no longer be any reason to suspect that the interaction of an economic and a non-economic form of marginalization would manifest in income at all.

Finally, the class and education interaction is interesting but perhaps not obviously meaningful. Since I did not have any reason to suspect that the interaction would be in one particular direction, only that an interaction might exist, only the interaction of class and the "less than high school education" dummy is significant in a two-tailed *p*-value context. Additionally, the *R*-squared increases less than a full percentage point.

Interestingly, when the non-significant interaction terms are dropped out, and only the "less than HS education + class" term is kept, the main effect of the "less than HS ed." dummy becomes statistically insignificant, although the interaction remains significant — in other words, for someone who has a zero on the class score, the income effect of having exactly 12 years of education is not different from the income effect of having a high school education, but for someone who has more than a zero on the class score, having less than a high school education is associated with less income than the same class score

[18] See *Black Feminist Thought*, 2nd ed. (New York: Routledge, 1990), p. 18.
[19] I reject the idea that class is simply a matter of "economics" but given the way in which I have had to operationalize the concept in this paper, it is not unreasonable to say that class is an "economic factor".

and a high school education — in fact, because the interaction coefficient is smaller than the class coefficient and because the "less than HS" dummy is statistically equal to zero, this might lead to the strange outcome that, for persons with less than a high school education, increases in class score are ultimately associated with decreases in income. As John Logan pointed out to me, the choice of models with similar *R*-squared values is to some degree a matter of "the story one wants to tell". Given how unintuitive this story is and the non-significant increase in explanatory power, I choose to simply drop the interaction terms from my model.

VI. Outliers, heteroskedasticity, and multicollinearity

Both logged-income and class were somewhat normally distributed, although both were still somewhat left-skewed with some outliers. Income has a couple of values which are less than zero on the logged scale (which means that the original values, i.e. an individual's income in thousands of dollars, must have been less than 1). This is reflected in the fact that about six, rather than five, percent of the data fell above the cutoffs when high values were isolated on Cook's d, DFFITS and DFBETAs (hat-values and Studentized residuals had means ≈<0.05; please see Stata output, section IV.i). In total, 84 cases (or about 11 percent of the data) scored above the cutoffs on one or more of the diagnostic scores. Examination of the individual outliers on each test revealed some "repeat offenders"; though it is hard to generalize, one might note that several of these cases, including three case IDs in a row (870-872), had exceptionally low incomes but reasonably normal scores on other variables.

Fortunately, when a regression without the 84 outliers was conducted following the procedure suggested by Rachel Gordon, there were no major changes to the parameter estimates or their relationship to one another. All of the main variables remained statistically significant but gender and class had slightly larger effects while all of the education variables had slightly lower effects. The "partial" and "full" models below capture, respectively, estimations without and with outliers included.

Heteroskedasticity-consistent standard errors are also shown in the table below under the column "robust standard errors". Predictably, there is some heteroskedasticity and so the standard errors increase when one accounts for this using the *vce(hc3)* command.

There were no serious problems with multicollinearity; the highest VIF was 1.49 (for "some

college"), which is well below the typical cutoff of four.

VII. Indirect effects and omitted variable biases

In this section, I consider some indirect effects, *not* for my preferred model, but for Model IV. This is for expository reasons. There are some relatively clear mediating effects of education and class on race in the lin-lin model worth exploring, but these become less clear in the log-lin model. I should also note that I "break a rule" of statistical analysis here by calculating indirect effects for a regression context with many predictors from a multivariate regression with only two predictors. As confirmed in conversation with John A. Logan, the correct calculation of indirect effects becomes significantly more complicated in the context of more than two predictors. So, I follow Prof. Logan's advice to simply "break this rule" and calculate indirect effects for the multi-predictor case from the two-predictor case. I also follow his suggestion to re-convert a polytomous categorical variable which has been "dummied up" into multiple binary variables (education) back to a single variable to show indirect effects.

I explore two effects in this section. First, I investigate whether the effect of race is mediated by education.[20] Second, I examine some omitted variable bias by scrutinizing my class variable more closely; I suggest that there are important dimensions of class that it leaves out.

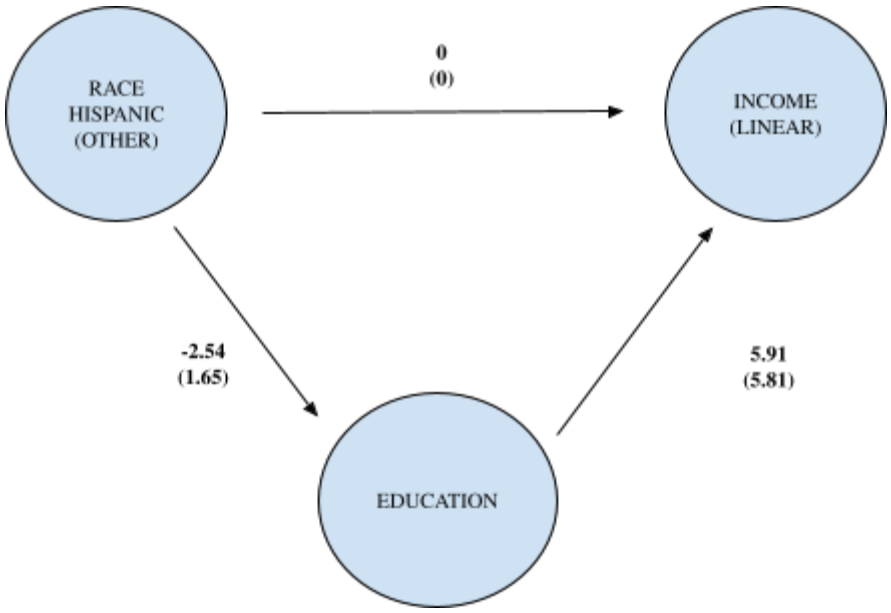
The mediating effect of education is somewhat difficult to capture. When a basic race and gender model is estimated (Model II), the addition of race (Model III) causes the effect of each race category to disappear, causing me to suspect that education mediates race. When bivariate income-on-race models are estimated, however, the addition of education to each model is only significant for the race-on-Hispanic and race-on-other models; the coefficient for education is insignificant in the race-on-Black model (please see Table 3 in the appendix for all of the regression results referenced in this section, and Figure 3 for the path diagram). N.b. that since the pathways are very similar, I have elected to put both in the same diagram, with Hispanic respondents on top and "other" respondents on the bottom in parentheses.[21] The

[20] I originally included an investigation of the possible mediating effect of class on education and the possibility that health and age might be confounders. This proved to simply take up too much space. Additionally, the missing data problem outlined in Appendix WHATEVER made this difficult.

[21] I am very strict in adhering to the 0.05 alpha cutoff here. Technically, a regression of income on "other race" and education generates a coefficient for the race dummy whose two-tailed *p*-value is 0.052. But, it would be unscientific to go back and use a *one*-sided *p*-value just because I now know that "other" has a strong positive association with income — when I started, I genuinely did not know which direction the association of "other" and income would be, so I continue to use a two-tailed *p*-value. My alpha is 0.05 and it would degrade the meaning of a

causality here is simple: Hispanic and "other" respondents have, respectively, less and more education on average; those with more education have higher incomes, on average (demonstrated in my Stata output, Part VII and shown in Figure X below).

Figure 3. Education mediates the effect of race on income



For Black respondents, this pathway is not valid: controlling just for education does not eliminate the statistical significance of the Black dummy coefficient because the average education of Black respondents is not statistically different from non-Black respondents (demonstrated in Stata output, Part VIII). Neither, however, does controlling for gender. But, if one controls for both class and gender, the direct effect of being Black on income becomes mathematically equivalent to zero. I do not show this causal pathway since it would seem, to me, to require structural equation modelling which is capacity. Also, regressions with categorical outcome variables follow their own rules which I do not yet understand. Suffice it to say that score on the class scale is positively associated with income and negatively associated with the Black dummy variable (see Stata output, Part VIII). The Black categorical variable is also positively associated with gender (i.e. being a woman), which is negatively associated with income.

This latter, mediating effect of gender on race needs explaining. On its face, this makes little sense: there is no reason, from the outset, to suspect that being Black somehow causes a respondent to be

---

cutoff to raise it.

more likely to be a woman. While gender is not necessarily determined at birth (only biological sex), there is no reason to suspect that Black respondents are more likely to undergo male-to-female gender conversions and consequently receive the lower "women's wage". In other words, it is unlikely that Blackness is a cause of gender. A more epistemologically-careful intuition, in my mind, would be that sex-at-birth rates and gender conversion rates are similar across races. So, controlling for gender might seem somewhat unintuitive. All of that said, however, it does make sense *mathematically* in this case: while the *overall* gender ratio favors women in both the full and slightly in the analytical sample (see above), the gender ratio for Black respondents even more sharply favors women — women are 62 percent of Black individuals in the full sample and 60 percent of the analytical sample. So, while there is no sense in which being Black "causes" one to be a woman, it is mathematically true in the sample that conditional probability of being a woman, given that one is Black, is higher (an independent group *t*-test confirms this, as shown in Stata output, Part VIII).

Now I can move on to consider the possible effects of omitted variables. In a sense, given how remote my measure of class is from the fundamental constituents of class in the Marxian sense, some class-related omitted variables are probably behind the relative lack of explanatory power that the class variable has (adding only one point to the *R*-squared when added to Model III above). Class might become a better predictor if we had information such as: 1) whether an individual makes most of her income from wages (as opposed to a salary, investment income, or cash transfers); and 2) a more precise formulation of the question — included in my class scale — concerning whether or not a respondent supervises anyone else.[22] My sense is that if we had a dummy that could indicate whether a respondent makes most of their money from wages and also a better dummy for whether someone is a supervisor or not, these would both enhance the explanatory power of class (whether they were included in the scale or

[22] The GSS question simply asks respondents to record whether they supervise anyone or not; while every single person in my analytical sample had a supervisor (!), almost 40 percent answered "yes" on the question *wksup*, which asks if the individuals or their spouses supervise anyone at work. This number is probably inflated because it allows for double counting (if both spouses answer the survey) but if we bracket that problem, this suggests something on the order of a 2:3 ratio of supervisors to workers in the US economy — perhaps this is a testament to some kind of fundamental decline in the US economy but it seems more likely that the question is just unclear (in my view, a more revealing question would ask about the respondent *only* and would in some way differentiate, say, managers at large banks from managers at fast food restaurants who, though they supervise others, are also on the proverbial shopfloor themselves).

as separate "class effects"); the linear association of class with income would increase, as would its ability to improve the explanation of variation in income.

VIII. Conclusions

My very concise interpretation of my final results is that my measure of class *does* positively contribute to the understanding of income variation. When compared to the "one-off" effect of being female, or the overall education effect (standardized), it is small, but it interesting and encouraging that the coefficient is at least in the right direction, so to speak. It is interesting that *none* of the race categories with which I originally began continue to retain their explanatory power in the final model, but I believe that the mediating effect of education explains the case of Hispanic and "other" respondents quite nicely, and it appears that, with some more experience under my belt, it would be fairly easy to calculate a simple structural equation model that showed how the effect of being Black is mediated by class (it would also be interesting to know more about the overrepresentation of women within the sample of Black respondents and how, if at all, that should be controlled for).

Some important limitations, as mentioned above, are: the relative (but not total) lack of objective information in my class scale; the imprecisions of the objective information that *does* exist; the trickiness of the "other" category; the conceptual question of whether it quite makes sense to say that class can really "explain variations in" income to begin with; the fact that I do not have a plausible mechanism by which education can directly account for variations in income.

IX. Appendix

TABLE 4. *Side-by-side model estimations showing indirect effects of various variables on logged income.*

Variables								
	Hisp. only	Hisp. + educ.	other-only	other + educ.	Black-only	Black + educ.	Black + gender	Black + gender + class
Hispanic	**-14.1*** **	1.05						
	(4.02)	(3.85)						
Other			**21.79*** **	12.19*				
			(6.87)	(6.28)				
Black					**-9.22** **	**-7.63** **	**-7.16***	-4.97
					(4.15)	**(3.76)**	**(4.09)**	(4.01)
Educ.		**5.95*** **		**5.81*** **		**5.879*** **		
		(0.46)		**(0.44)**		**(0.440)**		
Gender							**-16.71*** **	**-16.27** **
							(2.89)	**(2.83)**
Class								**8.541*** **
								(1.42)
Obsv.	797	797	797	797	797	797	797	797
R^2	0.015	0.184	0.012	0.188	0.006	0.189	0.046	0.009

Standard errors in parentheses.
Results are **bolded** when significant at $\alpha = 0.05$ *for the appropriate tail-test.*
* $p < 0.1$
** $p < 0.05$
*** $p < 0.01$

Source: GSS 2012.

TABLE 3. *Demonstration of certain selected interactions.*

Variables	Model IX (Gender x Ed)	Model X (Class x Gender)	Model XI (Class x Ed.)
Gender	**-0.426*** **	**-0.528*** **	**0.360*** **
	(0.124)	**(0.186)**	**0.0632)**
Less than HS educ.	**-0.417*** **	**-0.341*** **	0.393
	(0.154)	**(0.116)**	(0.290)
Some college	0.177	**0.201** **	0.462*
	(0.127)	**(0.0878)**	(0.257)
BA	**0.884*** **	0.833***	1.012***
	(0.137)	(0.0973)	(0.321)
Postgrad	**0.837*** **	0.952***	1.212***
	(0.131)	(0.0942)	(0.285)
Class	**0.123*** **	0.0899*	-0.0767
	(0.0321)	(0.0468)	(0.0856)
Gender x less than HS ed.	0.179		
	(0.232)		
Gender x some college	0.0603		
	(0.176)		
Gender x BA	-0.0938		
	(0.194)		
Gender x postgrad	0.237		
	(0.188)		
Gender x class		0.0596	
		(0.0633)	
Class x HS education			0.294***
			(0.107)
Class x some college			0.199*

			(0.106)
Class x BA			0.225*
			(0.120)
Class x postgrad			0.198*
			(0.111)
Observations	797	797	797
R-squared	0.257	0.255	0.261

Standard errors in parentheses.
Results are **bolded** when significant at $\alpha = 0.05$ *for the appropriate tail-test.*
* $p < 0.1$
** $p < 0.05$
*** $p < 0.01$

Source: GSS 2012.